# Введение

В рамках учебной практики была выполнена научно-исследовательская работа по теме «Вопросно-ответный поиск».

Согласно программе учебной практики направления подготовки 02.03.02 «Фундаментальная информатика и информационные технологии» целями практики являются:

* формирование профессиональных навыков в проведении научных исследований;
* формирование навыков использования современных научных методов для решения научных и практических задач;
* формирование практических навыков написания вспомогательных программных комплексов для проведения вычислительных экспериментов;
* формирование общекультурных, общепрофессиональный и профессиональных компетенций в соответствии с ОС ВО РУДН;
* формирование навыков оформления и представления результатов научного исследования;
* формирование навыков работы с источниками данных.

Там же определены задачи практики:

* формирование у студентов навыков в области изучения научной литературы и (или) научно-исследовательских проектов в соответствии с будущим профилем профессиональной деятельности и применения новых научных результатов;
* обучение правильному составлению научных обзоров и отчетов;
* формирование навыков решения конкретных научно-практических задач самостоятельно или в научном коллективе;
* обучение навыкам работы с прикладными комплексами программ для проведения вычислительных экспериментов;
* формирование способности разработки вспомогательных программных инструментов;
* обучение подготовке научных публикаций;
* формирование способности проводить научные исследования и получать новые научные и прикладные результаты.

Для достижения целей и решения поставленных задач в рамках учебной практики я выполнил обзор публикаций российских и международных научных изданий по теме практической работы, которая определена как «Вопросно-ответный поиск». Также был выполнен обзор датасетов, подходящих для данного исследования.

# Основная часть

В ходе данной работы я выполнил обзор публикаций российских и международных научных изданий по теме практической работы, которая определена как «Вопросно-ответный поиск».

## Обзор публикаций

### 1. Сопоставление синтаксических графов для улучшения модели вопросно-ответной системы [1]

В работе рассматривается один из методов улучшения модели вопросно-ответной системы – сопоставление синтаксических графов. Так как в настоящее время существует очень мало вопросно-ответных систем для русского языка, а разработанные на сегодняшний день системы часто выдают неудовлетворительные результаты, потому что дня него практически не используются семантические роли слов, а также системы способны отвечать только на простые работы, и в их работу заложены в основном алгоритмы статистического анализа текста, эффективность которых зачастую может оказаться недостаточной, становится очевидным факт, что необходимо повышать точность и быстродействие существующих систем .

Целью данной работы является разработка алгоритма, который позволяет автоматически находить ответы на вопросы к конкретным элементам текста. Исследование проводилось на текстах небольшого объема, состоящих из простых распространенных предложений на русском языке. Поиск ответа производился путем сравнения синтаксического графа вопроса, в котором вопросительное слово или словосочетание заменялось на маску, и синтаксических графов всех предложений текста. Наилучшие совпадения маски со словом или словосочетанием в тексте, отсортированные по релевантности, выдаются как ответ на вопрос. Разработанный алгоритм показал как высокое быстродействие (на предварительно размеченных текстах), так и высокую точность ответов.

Ввиду роста актуальности темы было необходимо разработать модель, с возможностью расширять вопросно-ответную систему, позволяющую дать точный ответ на поставленный вопрос, и имеющую высокое быстродействие. Расширяемость системы достигается за счет разделения системы на две независимых подсистемы, каждая из которых может быть улучшена, или заменена.

Моделью предложения является синтаксическое дерево, полученное в ходе автоматического анализа. С помощью морфологического анализа каждому слову присваивается грамматический признак, и с помощью определенного набора правил составляется дерево вероятности подчинения слов. Следующим шагом вероятности корректируются, путем учитывания контекста предложения, и замены местоимений антецедентами, процесс может повторяться несколько раз, пока не будет получен дерево с наибольшей вероятностью. Полученные синтаксические дерева позволяют проводить различные операции с предложениями, для выдачи максимально точного ответа.

В результате анализа модели система выдает, в виде конкретного слова или словосочетания, ответ на заданный вопрос. В ходе эксперимента, проведенного в работе, была показана высокая устойчивость алгоритмов, не менее 90%, в случае правильно разобранного вопроса и предложения, а в случае некачественного анализа – менее 35%. Это показывает, что решающим фактором верного ответа является качество автоматического анализа. Проведенные эксперименты помогли определить пути совершенствования системы

### 2. Диалоговый агент с обучаемым диалоговым менеджером [2]

В данной работе рассматривается создание диалогового агента в рамках соревнования по разработке диалоговых агентов. Разработанные ранее системы использовали подход, основанный на правилах для разработки стратегии ведения диалога. Такие системы трудны в поддержке, и требуют написания правила под каждую тематику диалога, также такие системы обладают низким качеством ответа. Соревнования по разработке диалоговых агентов общего назначения являются одним из путей привлечения людей к оценке качества, для создания более совершенных систем.

В данной работе была поставлена цель научить бота приветствовать пользователя, отвечать на вопросы о тексте, а также задавать их. Так как заданный текст может быть длинным, бот должен уметь автоматически реферировать его.

Часто пользователь имеет желание поговорить с ботом на общие темы, поэтому он также должен уметь это делать. Для поддержания таких диалогов были разработаны три навыка разговоров на общие темы. Разговор о тексте также может включать в себя вопросно-ответный сценарий, для чего были разработаны навыки генерации вопроса, ответа на вопрос, и проверки ответа на заданный вопрос. Бот был обучен на основе датасета SQuAD с помощью библиотеки OpenNMT. Для автоматического реферирования текста был также разработан соответствующий навык. Этот навык может помочь пользователю сэкономить время на вовлечение в беседу. Навык работает на основе датасета Gigaword на OpenNMT. При упоминании основной темы текста навык определения темы текста помогает с вовлечением пользователя в беседу. Навык определения был реализован с помощью библиотеки для тематического моделирования BigARTM. Дополнительно были реализованы более простые навыки, основанные на случайной отправке заготовленных фраз – навык приветствия и навык общих фраз.

Диалоговый менеджер состоит из двух классификаторов, которые принимают на вход контекст диалога или реплику пользователя, а на выход возвращают навык, выдающий ответ. Первый классификатор использует небольшую обучающую выборку, состоящую из нескольких фраз на навык, а второй – использует большую обучающую выборку, состоящую из нескольких фраз на навык и библиотеки fastText.

Анализ результатов показал, что наиболее популярным навыком стал разговорный, следовательно этот навык необходимо развивать, но несмотря на то, что навык реферирования вызывается всего один раз, это не значит, что он бесполезен, потому что этот навык вовлекает пользователя в беседу. Наличие нескольких разговорных навыков позволяет боту выдавать качественные реплики, и вовлекать пользователя в диалог.

### 3. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus [3]

Статья рассматривает отличительные особенности вопросно-ответного поиска. Показаны базовые принципы работы интеллектуальной поисковой системы Exactus. Вопросно-ответный поиск имеет существенные отличия от поиска, по ключевым словам, или его аналогов. Вопросы для поисковых систем чаще всего формируются в виде вопросительного предложения. Основной особенностью вопросно-ответного поиска заключается в поиске слова, словосочетания, или целого предложения, которое не содержится в явном виде в запросе. Вопросно-ответный поиск имеет другую цель в сравнении с обычным поиском. Его задача – найти фрагмент документа (веб страницы), содержащий точный ответ на заданный вопрос. Статья описывает ситуативно-реляционную модель текста, описывающую его семантику.

Если обратиться к теории коммуникативной грамматики русского языка, которая описывает подход к описанию русского синтаксиса, то можно выделить идею, заключающуюся в том, что синтаксис должен изучать именно осмысленную речь, а синтаксические правила должны учитывать категориальные значения слов, чтобы иметь возможность определять обобщенный смысл любой синтаксической конструкции.

Для построения модели предложения необходимо провести морфологический, синтаксический и семантический анализ текста. Для вопросно-ответного поиска анализ запроса, и анализ документов коллекции текстов. По тексту, для каждого предложения, строится семантическое дерево, которое является образом запроса и документов. Для реализации поиска необходимо вычислить релевантность результатов. Приоритетно именно сравнение синтаксем и связей, а не совпадение словоформ запроса.

В результате работы с помощью независимых экспертов удалось проверить работоспособность поисковых алгоритмов для дальнейшего развития системы.

### 4. Взвешенная погрешность — новая метрика для оценки качества валидации ответов в задаче вопросно-ответного поиска [4]

В данной работе рассмотрены существующие подходы к экспериментальной оценке качества вопросно-ответных систем, в частности модуля валидации ответов. Обоснован выбор новой метрики для выполнения экспериментов. Для оценки вопросно-ответной системы применяются некоторые метрики. Чтобы оценить валидацию ответа можно сравнить проходы системы в разных конфигурациях. Сравнив результаты этих прогонов, можно оценить вклад предлагаемой реализации модуля валидации ответов в качество вопросно-ответной системы. Процедура оценки вопросно-ответной системы очень трудоемка, так как требует работы нескольких асессоров, оценивающих результаты прогонов, но она позволяет выявить наиболее точные алгоритмы.

По результатам участия кампании РОМИП 2010, было принято решение исследовать подзадачу валидации ответа как задачу бинарной классификации. Была предложена новая метрика — взвешенная погрешность Eα, в отличие от традиционной F-меры учитывающая исходы true-negative, являющиеся важной категорией ответов для задачи валидации ответов.

5. Семантическая технология проектирования интеллектуальных вопросно-ответный систем [5]

Объектом рассмотрения в данной работе являются интеллектуальные вопросно-ответные системы, которые дают ответы на широкий спектр вопросов. Предложена модель данного класса систем, а также семантическая технология их проектирования. Поиск информации отнимает много времени, потому что существующие поисковые системы выдают большое количество ссылок, которые зачастую не имеют прямого отношения к теме.

Существующие подходы оптимизации информационного поиска в настоящее время лежат в плоскости разработки вопросно-ответных систем, в которых осуществляется сопоставление вопросов пользователей с требуемой информацией. Существующие системы AllQuest и AskNet Global Search ориентированы только на анализ и выявление семантических отношений между объектами предметной области в проиндексированных текстах. Данное обстоятельство накладывает следующие ограничения: нет возможности строго формально установить семантические отношения между объектами в тексте; невозможно сгенерировать ответ пользователю, когда такого ответа нет в проиндексированных текстах (т.е. в текущем информационном состоянии системы); не поддерживаются вопросы на выявление соответствий и аналогий между объектами и понятиями. Для устранения перечисленных ограничений требуется создание следующего поколения вопросно-ответных систем.

В работе были описаны основные понятия и принципы построения интеллектуальных вопросно-ответных систем. Прицнип многократного использования компонентов системы позволяет сократить сроки проектирования подобных систем за счет возможности использовать уже готовые типовые фрагменты баз знаний, операций и самих машин обработки знаний, пользовательских интерфейсов, оформленных в видекомпонентов интеллектуальной собственности.

Также в работе был описан язык запросов, который относится к семейству совместимых семантических языков. Язык запросов предназначен для формального описания поискового предписания вопросно-ответных систем, с целью удовлетворения информационной потребности пользователя. Объектами анализа языка вопросов являются классы вопросов в соответствии с семантической типологией вопросов.

Предложенная в работе семантическая технология предназначена для проектирования класса интеллектуальных вопросно-ответных систем, использующих в качестве формальной основы универсальный семантический код. Наличие совместимых семантических технологий проектирования баз знаний, машин обработки знаний и пользовательского интерфейса позволяют сократить сроки проектирования рассматриваемого класса систем за счет унификации способов представления знаний и легкой интегрируемости указанных компонентов.

### 6. Исследования и обзоры в вопросно-ответной системе (Research and reviews in question answering system) [6]

Вопросно-ответные системы – это автоматизированный подход к получению верных ответов на вопросы, заданные человеком на естественном языке. В данной работе рассматривается таксономия для характеристики вопросно-ответной системы, также кратко рассматриваются основные системы контроля качества ответов, и было проведено сравнение подходов, основанных на определенных особенностях системы контроля качества.

Главной проблемой для существующих систем контроля качества является правильное понимание вопросов на естественном языке, для выдачи верных ответов. Задача «ответ на вопрос» включает в себя методы искусственного интеллекта, обработки естественного языка, статистического анализа, сопоставления шаблонов, поиска и извлечения информации.

Были рассмотрены следующие подходы: лингвистический, статистический, составление шаблонов, на основе поверхностного шаблона, на основе шаблона. Все эти подходы хорошо работают в своих областях. Например, системы контроля качества, основанные на лингвистическом подходе, в основном, были построены на основе базы знаний для конкретной области. Статистический подход полезен для большого количества данных, система сможет давать ответы на сложные вопросы. Шаблонный подход вместо критического лингвистического анализа использует красноречие текстовых шаблонов. Даже несмотря на некоторые сбои, это удивительно эффективный метод, если использовать интернет, как источник данных. Однако, данному методу не хватает семантического понимания и рассуждений.

Так как все рассмотренные методы работают хорошо только в своих областях, стала очевидна необходимость создания универсального метода с гибридным подходом поиска. Были рассмотрены различные методы, такие как: MULDER, ASQA и IBM WATSON QA System.

В ходе работы получен вывод, что гибридные подходы обеспечивают лучшие результаты, но тем не менее очевидно, что в центре внимания останутся простые методы вопросно-ответных систем, потому что гибридные техники очень специфичны в работе.

### 7. Нейронная вопросно – ответная система для основных вопросов о подпрограммах (A Neural Question Answering System for Basic Questions about Subroutines) [7]

В работе рассматривается система вопросно – ответного поиска для базовых вопросов о подпрограммах. Цель работы сделать шаг к полноценной системе контроля качества для инженеров-программистов, необходимо продемонстрировать, как адаптировать эти технологии, чтобы можно было извлечь информацию из огромных наборов данных. Конечно, невозможно создать систему, которая будет отвечать на любые вопросы о программировании. Но идея состоит в том, чтобы создать систему, способную давать машинные ответы на информационные потребности программистов, которые настроены на контекст программного обеспечения и отдельные вопросы программистов. Для решения этой задачи необходимо решить несколько проблем, которые являются барьером в дальнейшем прогрессе. Успешная система должна не только ответить на узкую проблему, но и дать представление о том, как моделировать и извлекать функции из исходного кода, как интерпретировать информационные потребности программиста и как понимать словарь, который используют программисты, который отличается от общего использования слов.

Далее рассматривается справочная информация об интерактивных диалоговых системах и нейронном энкодере-декодере. Определяются методы выполнения работы. В ходе выполнения был создан набор данных для обучения, который позволил системе обучиться методам языка Java из более чем 10 тыс. проектов.

Затем была проведена оценка результатов, для которого были приглашены программисты. В ходе эксперимента выявлено, что модель может достоверно отвечать на поставленный вопрос, также стоит отметить, что система не предназначена для использования сама по себе, а предназначена в качестве компонента гораздо более широкой системы диалога.

### 8. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions [8]

Так как вопросно-ответные системы в биомедицинской области полноценно могут отвечать лишь только на ограниченное количество вопросов, а также требуют доработок для увеличения точности, и производительности. В работе была представлена семантическая биомедицинская система контроля качества под названием SemBioNLQA. Эта система способна обрабатывать запросы разных типов, на естественном языке.

Со временем усвоение информации в области биомедицины становится все сложнее из-за постоянно растущего объема, поэтому для поиска нужной информации может потребоваться много времени. Поэтому все чаще используются интеллектуальные методы анализа текста, однако в существующих системах размер ответа может быть велик, что также осложняет поиски точной информации, а некоторые вопросы вообще остаются без ответа. Поэтому важно иметь возможность получить отчет на любой тип вопроса на естественном языке. Вопросно-ответные системы, нацелены на предоставление точных ответов, путем автоматического анализа тысяч статей за несколько секунд. Поэтому целью данной работы является описать процесс проектирования и развития системы SemBioNLQA. Преимущества этой системы в том, что она решает проблемы предыдущих систем, такие как: Невозможность предоставить ответ для разных типов вопросов, невозможность предоставить идеальный ответ на естественном языке, проведение экспериментов с огромным количеством данных.

В работе были описаны методы, которые помогли достичь таких высоких показателей качества ответов. В качестве основных методов использовались: классификация вопросов и переформулирование запросов, извлечение информации из документов, поисковый проход, извлечение вопроса, вопрос да/нет, фактоидные вопросы, задание вопросов, обобщенные вопросы.

Результаты ручной оценки показали, что SemBioNLQA достигла лучших ответов, в сравнении с самой современной системой, а также смогла обеспечить конкурентоспособную альтернативу ей, помогая достичь точных ответов. Также, в будущем, рассматривается повышение производительности системы, в качестве основных направлений можно выделить: разные способы глубокого обучения, увеличение пар вопрос-ответ, улучшение работы с источниками.

### 9. Automatic question-answer pairs generation and question similarity mechanism in question answering system [9]

Вопросно-ответные системы играют важную роль в получении автоматических ответов на вопросы, с помощью информационных баз знаний. Данная работа включает в себя описание генерации, понимания и ответов на вопросы, преодолевая все ограничения систем. Первые вопросно-ответные системы появились еще в 1960-х годах, с того времени технологии поменялись, а количество информации возросло, поэтому доверие к системам повысилось. Системы научились выдавать ответ, максимально близкий к естественной речи. Но есть в этих системах и ограничения – это вопросы, на которые невозможно ответить. Если вопрос неверен, но связан с контекстом, то система может задать уточняющие вопросы. Точность и скорость ответов на вопросы от систем значительно превосходит человеческую.

Данная работа вносит определенный вклад в развитие вопросно-ответных систем, такие как:

Автоматическая генерация возможных пар вопрос-ответ с учетом отрывка

Ввод механизма подобия вопросов, на которые нет ответа, а также нерелевантные вопросы.

Объединение системы генерации вопросов с системой ответов на вопросы, создав приложение «Автоматическая система генерации пар вопрос-ответ»

Автоматическая система генерации пар вопрос-ответ генерирует все возможные вопросы, и ответы на них. Такая система может применяться в разных областях. За счет введения механизма подобия вопросов система имитирует человеческие рассуждения, чтобы определить есть ли ответ на вопрос, существующие системы не умеют делать этого. Если вопрос не имеет ответа, то он не передается в вопросно-ответную систему. Также такая система не требует крупных вычислительных мощностей, поскольку в ней нет процесса обучения. Данный механизм можно включать в современные системы, за счет него модели могут сфокусироваться на ответах на вопрос, что повышает их производительность. А автоматически сгенерированные вопросно-ответные пары можно использовать, как набор данных для обучения вопросно-ответных систем.

### 10. Bangla Intelligence Question Answering System Based on Mathematics and Statistics

В работе рассматривается информационная система ответов на вопросы Bangla. Каждый хочет иметь систему, отвечающую на вопросы на своем родном языке, чтобы вопросы не ограничивались никакими правилами, или даже областью знаний. Новый подход соответствия потребностям пользователя заключается в проведении фактического анализа вопроса с лингвистической точки зрения и попытке понять, что на самом деле имеет в виду пользователь.

Система состоит из трех основных модулей: сбор данных, обработка информации и вопросов пользователей, а также установление взаимосвязи между ними. В процессе работы были внедрены математические и статические процедуры поиска информации, для предварительной обработки данных применили алгоритмы Хоббса, Edit Distance, и т.д., для упрощения предварительной обработки был разработан модуль, содержащий математические методы. Для разработки системы оценки качества были использовались пять основных компонентов, таких как лингвистическая предварительная обработка, распознавание сущностей, сопоставление элементов онтологии, обнаружение семантических связей, а также формулировка пар вопрос-ответ. Отличие от остальных систем состоит в том, что система работает на основе математики и статистики.

В процессе оценки исследований было показано, что разработанный чат-бот куда более современный, чем другие существующие. В будущем такая система может быть включена в образовательные, промышленные, деловые сферы, а также для личных задач и голосовых ассистентов. Дальнейшее развитие возможно с помощью алгоритмов глубокого обучения на рекуррентных нейронных сетях.

## Обзор датасетов

### Question-Answer Dataset

В данном датасете имеется все, что нужно для того, чтобы опробовать свои силы в задаче вопросно-ответного поиска. Датасет содержит 3 файла с вопросами, а также почти 700 000 слов очищенного текста с Википедии, с помощью которого были созданы вопросы. Также имеется файл вопрос-ответ, который содержит пары вопрос-ответ, название статьи в Википедии, оценка

сложности вопроса, рейтинг сложности и имя файла, содержащего статью. Также из данного датасета были исключены «плохие» вопросы.

### Stanford Question Answering Dataset

Датасет, созданный Стэнфордским университетом. Этот набор создан для решения задачи понимания прочитанного текста. Датасет также содержит вопросы, составленные на основе более 500 статей из Википедии, а ответом на каждый вопрос является сегмент текста или словосочетание. База содержит около 100 000 пар вопрос-ответ.

### SberQuad

Датасет, созданный российской компанией Сбер. Этот набор данных создан для понимания текста, состоит из вопросов, заданных краудворкерами на основе статей с Википедии, где ответом на вопрос является фрагмент текста. Это российский аналог Stanford Question Answering Dataset, но он не был должным образом представлен научному сообществу, поэтому на его основе датасета существует очень мало исследований.

### DaNetQA

Датасет DaNetQA – это набор данных, созданный для решения задачи бинарной классификации, т.е. ответов на вопросы, с ответом «да» или «нет». Вопросы возникают естественным образом, что значит, что генерируются без ограничений. Набор состоит из трех файлов – test, train, val. В каждом файле находится набор данных в форме (вопрос, созданный человеком/данный, отвечающие на него/ответ «да» или «нет»). Данные были собраны из открытой энциклопедии – Википедия.

Данный датасет является частью проекта Russian SuperGLUE, который создан с целью научить нейросети работать с русским языком, а также помогает решить любые задачи обработки естественного языка, выбирать наиболее оптимальные метрики для решения конкретных задач. Также Russian SuperGLUE является рейтингом NLP-систем.

### RWSD

Также, как и предыдущий датасет, RWSD является частью российского проекта Russian SuperGLUE. Данный датасет направлен на обучение модели решать задачу различения двусмысленности предложений. Ответом на вопрос также является простой ответ «да» или «нет», но он отличается от датасета DaNetQA тем, что данными для обучения являются два предложения, которые незначительно отличаются, но имеют разный смысл. Задача модели – сделать логический вывод, и ответить на поставленный вопрос верно. Для данной задачи не подходит классическое ориентирование на порядок слов и структуру предложения, так как неоднозначность предложения не может быть устранена данными способами.

# Заключение

За период практики, которую я проходил на кафедре ИТ факультета ФМиЕН РУДН, были достигнуты все цели и решены все поставленные задачи, определенные в программе учебной практики направления подготовки 02.03.02 «Фундаментальная информатика и информационные технологии» (см. введение отчета по практике).

При прохождении практики я разобрался с научной терминологией области исследований; научился осуществлять сбор, анализ и обработку данных, необходимых для решения профессиональных задач; собирать, обрабатывать и интерпретировать данные современных научных исследований, необходимые для формирования выводов по соответствующим научным исследованиям; осуществлять целенаправленный поиск информации о новейших научных и технологических достижениях в сети Интернет и из других источников; оформлять результаты своей работы.

В результате прохождения данной практики я приобрел следующие практические навыки, умения, универсальные и профессиональные компетенции:

* способностью к самоорганизации и к самообразованию;
* способностью решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности;
* способностью к самостоятельной научно-исследовательской работе;
* способностью к определению общих форм и закономерностей отдельной предметной области;
* способностью строго доказать утверждение, сформулировать результат, увидеть следствия полученного результата;
* способностью публично представлять собственные и известные научные результаты;

Выполненный во время проведения учебной практики обзор публикаций научных изданий по теме Вопросно-ответный поиск позволит мне обосновать актуальность выбранной темы научно-исследовательской работы, а также более полно раскрыть постановку задачи и методы исследования при ее выполнении.